交叉熵正则化的数学解释

何沧平
cangping@staff.weibo.com
cphe@lsec.cc.ac.cn
微博

摘　要

本文通过严格数学分析找出了交叉熵过拟合的成因：边界样本的损失贡献比重大且随法向量增长而加速增大、边界样本分布散乱，顺便理清了正则项的作用机理。

关键词: 交叉熵, 深度学习过拟合, 正则化

Mathematical mechanism of regularization of cross entropy*

He Cangping
cangping@staff.weibo.com
cphe@lsec.cc.ac.cn
WEIBO.COM

Abstract

In this paper, I found the two reasons of overfitting of cross entropy: boundary samples occupy a larger and larger share as the length of normal vector becomes longer and longer, boundary samples do not fit their probability density function well.

Keywords: cross entropy, deeplearning overfitting, regularization

1　引言

交叉熵是机器学习的一个常用损失函数，可以用于简单的算法，例如逻辑回归 [14]，也可以用于复杂的模型，例如 BERT[3]、SimCLR[2]。

交叉熵有一个无法完美解释的的问题，过拟合现象，即训练一段时间以后，随着训练样本集上的正确率逐渐提高，测试样本上的正确率却不再提高甚至反而下降。过拟合的根本原因尚无共识，目前的应对办法是在损失函数中添加正则化项 [11]，阻止参数变得过大，至于多大算是“过大”，没有具体定义。

*完稿日期：2022 年 11 月 26 日

通常的解释是模型过于复杂 [11, 14]，要用相对简单的模型来缓解过拟合现象；至于过拟合的成因，可用“偏差-方差分解”[5, 14] 来解释，[4] 还讨论了过拟合与噪声、多重假设检验的关系。缓解过拟合的常用手段是添加正则化项，[11] 对比了 L_1 正则化和 L_2 正则化的特点。

虽然正则化缓解了过拟合现象，但它带来了新的麻烦：正则化系数的选择缺少理论指导，只能针对具体训练样本多次试探；正则化还增加了模型复杂度，求解最优化问题需要大量的技巧 [1, 6–10, 12, 13]。

本文跳出常规的概率视角，通过观察直观的几何图像，用严格的数学公式证明，过拟合的原因有两个：边界样本的损失贡献比重大且随法向量增长而加速增大、边界样本分布散乱；虽然法向量过大只是过拟合的表象，但是控制法向量模长却能够切实缓解过拟合，因此各种正则化手段有效。

本文后续内容这样组织。第 2 节定义几个函数，为后文公式推导做准备；第 3 节给出交叉熵的公式；第 4 节给出过似和实例，证明过拟合成因；第 5 节证明 2 个定理，法向无限和法向有界。

2 符号定义

作为准备，本节定义几个函数。目前流行深度学习软件包中，例如 pytorch，数组的组织方式是行优先，因此本文中的向量、矩阵也按行优先来定义。

任意给定正整数 m 和 d，行向量用黑体小写字母表示，形式为 $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$。矩阵用大写字母表示，形式为

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{md} \end{bmatrix}.$$

软大函数 (softmax) 定义为

$$\begin{aligned} \mathrm{smax}(\boldsymbol{x}) &= \frac{1}{\sum_{i=1}^{i=n} e^{x_i}} (e^{x_1}, e^{x_2}, \ldots, e^{x_d}), \\ \mathrm{smax}(X) &= \begin{bmatrix} \mathrm{smax}(x_{1:}) \\ \mathrm{smax}(x_{2:}) \\ \vdots \\ \mathrm{smax}(x_{m:}) \end{bmatrix} = (\mathrm{smax}(x_{1:}); \mathrm{smax}(x_{2:}); \ldots; \mathrm{smax}(x_{m:})), \end{aligned}$$

这里的 $x_{i:} = (x_{i1}, x_{i2}, \ldots, x_{id})$，圆括号里的分号表示换行。

对向量或矩阵求对数时，对数作用到它们的每一个元素上，即

$$\begin{aligned} \log(\boldsymbol{x}) &= (\log(x_1), \log(x_2), \ldots, \log(x_d)), \\ \log(X) &= \begin{bmatrix} \log(x_{11}) & \log(x_{12}) & \cdots & \log(x_{1d}) \\ \log(x_{21}) & \log(x_{22}) & \cdots & \log(x_{2d}) \\ \vdots & \vdots & & \vdots \\ \log(x_{m1}) & \log(x_{m2}) & \cdots & \log(x_{md}) \end{bmatrix}. \end{aligned}$$

假设行向量 $\boldsymbol{b}=(b_1,b_2,\ldots,b_d)$，将行向量与矩阵相加定义为逐行相加，即

$$X+\boldsymbol{b}=\begin{bmatrix} x_{11}+b_1 & x_{12}+b_2 & \cdots & x_{1d}+b_d \\ x_{21}+b_1 & x_{22}+b_2 & \cdots & x_{2d}+b_d \\ \vdots & \vdots & & \vdots \\ x_{m1}+b_1 & x_{m2}+b_2 & \cdots & x_{md}+b_d \end{bmatrix}.$$

3 交叉熵

给定数据集 $D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\ldots,(\boldsymbol{x}_m,y_m)\}$, d 维行向量 $\boldsymbol{x}_i=(x_{i1},x_{i2},\ldots,x_{id})$，整数 $y_i\in\{1,2,\ldots,a\}$ 表示向量 $\boldsymbol{x}_i$ 归属的类，正整数 a 表示类别数量。

令 a 维向量 $\boldsymbol{\epsilon}_y=(0,\ldots,0,1,0,\ldots,0)$，即第 y 个元素为 1，其它元素均为 0。对任意给定的 $(\boldsymbol{x},y)$，令 $\boldsymbol{u}=\boldsymbol{x}W+\boldsymbol{b}$，这里的 W 是 $d\times a$ 矩阵，$\boldsymbol{b}$ 是长度为 a 的行向量。令

$$\boldsymbol{s}=\mathrm{smax}(\boldsymbol{u}) \tag{1}$$

显然 $\boldsymbol{s}$ 是长度为 a 的行向量。样本 $(\boldsymbol{x},y)$ 上的交叉熵为

$$l(x,y)=-\ln(\boldsymbol{\epsilon}_y\boldsymbol{s}^T).$$

因此，样本集 D 上的损失函数为

$$L(W,\boldsymbol{b})=\frac{1}{m}\sum_{i=1}^{m}l(x_i,y_i), \tag{2}$$

求解它的最小值

$$\{\hat{W},\hat{\boldsymbol{b}}\}=\arg\min_{W,\boldsymbol{b}}\frac{1}{m}\sum_{i=1}^{m}l(x_i,y_i), \tag{3}$$

得到最优参数 $\hat{W}$ 和 $\hat{\boldsymbol{b}}$。将最优参数代入式 (1)，就能预测任意样本 $\boldsymbol{x}$ 归属哪一类，即 $\arg\max_{j=1}^{a}s_j$。

3.1 二分类交叉熵

当 $a=2$ 时，矩阵 W 的元素记为

$$W=\begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ \vdots & \vdots \\ x_{d1} & x_{d2} \end{bmatrix},$$

列向量 $\boldsymbol{w}_1=(w_{11},w_{12},\ldots,w_{1d})^T$，列向量 $\boldsymbol{w}_2=(w_{21},w_{22},\ldots,w_{2d})^T$，列向量 $\boldsymbol{w}=\boldsymbol{w}_1-\boldsymbol{w}_2$。行向量 $\boldsymbol{b}$ 的元素记为 $\boldsymbol{b}=(\tilde{b}_1,\tilde{b}_2)$，$b=\tilde{b}_1-\tilde{b}_2$；行向量 $\boldsymbol{u}$ 的元素记为 $\boldsymbol{u}=(u_1,u_2)$，$u_1=\boldsymbol{x}\boldsymbol{w}_1+\tilde{b}_1$，$u_2=\boldsymbol{x}\boldsymbol{w}_2+\tilde{b}_2$；行向量 $\boldsymbol{s}$ 的元素记为 $\boldsymbol{s}=(s_1,s_2)$。对 $(\boldsymbol{x},y)$，如果 $y=1$，称 $\boldsymbol{x}$ 为正样本，此时 $\boldsymbol{\epsilon}_y=(1,0)$；如果 $y=2$，称 $\boldsymbol{x}$ 为负样本，此时 $\boldsymbol{\epsilon}_y=(0,1)$。

由式 (1) 得到

$$\begin{aligned} s_1 &= \frac{e^{u_1}}{e^{u_1}+e^{u_2}}=\frac{1}{1+e^{u_2-u_1}}=\frac{1}{1+e^{-z}}=\sigma(z), \\ s_2 &= \frac{e^{u_2}}{e^{u_1}+e^{u_2}}=1-\sigma(z), \end{aligned}$$

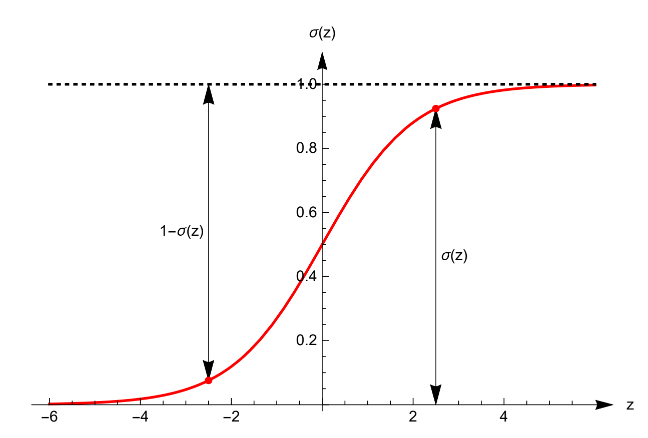


图 1: 二分类交叉熵：近似值 $\sigma(z)$(红色曲线) 与 $h(z,y)$(带双向箭头的直线)。

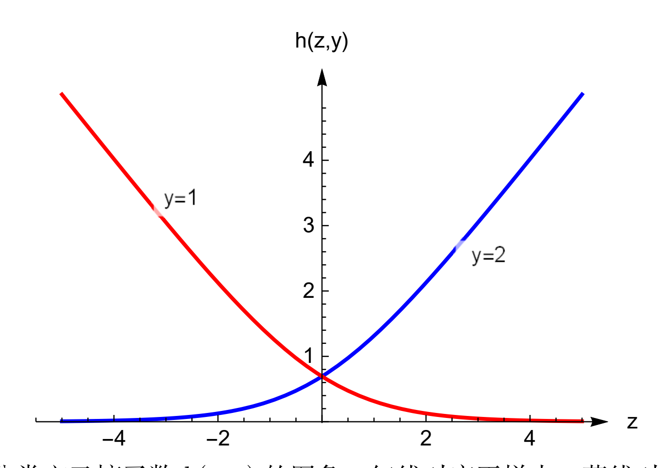


图 2: 二分类交叉熵函数 $h(z,y)$ 的图象。红线对应正样本，蓝线对应负样本。

这里的实数 $z = u_1 - u_2 = \boldsymbol{xw} + b$，Sigmoid 函数 $\sigma(z) = \frac{1}{1+e^{-z}}$。从而二分类交叉熵可以改写为

$$h(z,y) = \begin{cases} -\ln(\sigma(z)), & \text{如果 } y = 1, \\ -\ln(1-\sigma(z)), & \text{如果 } y = 2. \end{cases} \tag{4}$$

相应地，样本集 D 上的损失函数式 (2) 改写为

$$L(W,\boldsymbol{b}) = \frac{1}{m}\sum_{i=1}^{m} h(z_i,y_i), \tag{5}$$

最优化问题 (3) 改写为

$$\{\hat{W},\hat{\boldsymbol{b}}\} = \arg\min_{W,\boldsymbol{b}} \frac{1}{m}\sum_{i=1}^{m} h(z_i,y_i). \tag{6}$$

交叉熵用作损失函数，对任意 (z,y)，如果类别 y 正确，那么交叉熵 $h(z,y)$ 应该为 0 或者十分接近 0，从而要求 $\sigma(z)$ 对正样本趋向于 1，$1-\sigma(z)$ 对负样本趋向于 1。如图1 所示，红色曲线是 Sigmoid 函数 $\sigma(z)$；在 $\boldsymbol{x}$ 为正样本即 $y=1$ 时，用右侧双向箭头标记的距离反映曲线 $\sigma(z)$ 与 1 之间距离，z 越大，$\sigma(z)$ 越接近于 1；在 $\boldsymbol{x}$ 为负样本即 $y=2$ 时，用左侧双向箭头标记的距离反映曲线 $\sigma(z)$ 与 1 之间距离，z 越小，$1-\sigma(z)$ 越接近于 1。

图2画出了单个正样本 (红色) 和单个负样本（蓝色）的损失曲线。直观地理解，如果样本集是线性可分的，那么正样本对应的 z 越大，该样本上的损失函数值越小；负样本对应的 z 越小，

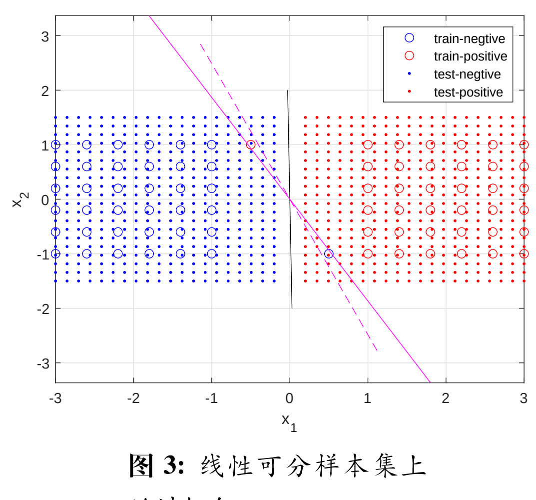


图 3: 线性可分样本集上的过拟合

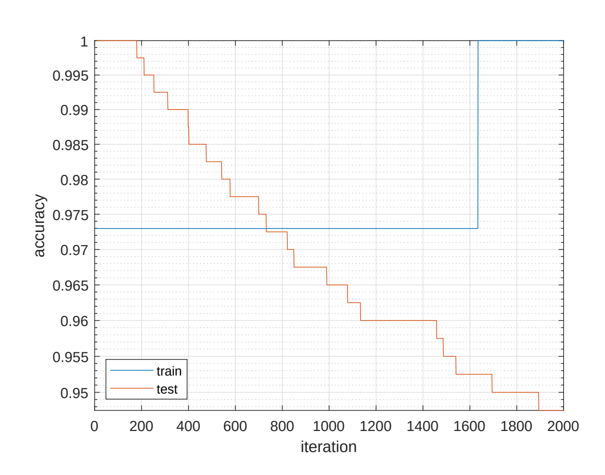


图 4: 线性可分样本集上的正确率

该样本上的损失函数值越小。从而，式 (3) 的计算结果是负样本向 z 负无穷方向移动，正样本向 z 正无穷方向移动，达到了分类的目的。

4 过拟合实例与成因

如果存在列向量 $\tilde{\boldsymbol{w}} = (\tilde{w}_1, \tilde{w}_2, \ldots \tilde{w}_d)^T$ 和行向量 $\tilde{\boldsymbol{c}} = (\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_d)$，且 $|\tilde{\boldsymbol{w}}| \neq 0$，使得对 $\forall(\boldsymbol{x}_i, y_i) \in D$ 有

$$\begin{cases} y_i = 1, & \text{如果 } (\boldsymbol{x}_i - \tilde{\boldsymbol{c}})\tilde{\boldsymbol{w}} \geq 0, \\ y_i = 2, & \text{如果 } (\boldsymbol{x}_i - \tilde{\boldsymbol{c}})\tilde{\boldsymbol{w}} < 0, \end{cases} \tag{7}$$

那么称数据集 D 是线性可分的，称 d 维平面

$$(\boldsymbol{x} - \tilde{\boldsymbol{c}})\tilde{\boldsymbol{w}} = 0 \tag{8}$$

为分隔面。显然，分隔面不唯一。

以交叉熵作为损失函数训练时，正确率通常会随着训练步数的增加而升高。有时在训练若干步以后，随着训练集样本上的正确率逐渐提高，测试集上的正确率不再提高甚至下降，这种现象称为过拟合。

为直观说明过拟合的成因，先给出 2 个没有实际意义的例子，它们分别对应线性可分的样本集和线性不可分的样本集。

4.1 线性可分样本集上的过拟合

图3中，蓝色圆圈是训练集中的负样本，红色圆圈是训练集中的正样本。训练集中的 36 个负样本均匀分布在区域 $[-3, -1] \times [-1, 1]$ 中，一个偏离主体的训练集负样本是点 (0.5,-1)。训练集中的 36 个正样本均匀分布在区域 $[1, 3] \times [-1, 1]$ 中，一个偏离主体的训练集正样本是点 (-0.5,1)，容易验证这个训练集线性可分。20×20 个蓝色小圆点是测试集中的负样本，它们均匀分布在区间 $[-3, -0.2] \times [-1.5, 1.5]$ 中；20×20 个红点小圆点是测试集中的正样本，它们均匀分布在区间 $[0.2, 3] \times [-1.5, 1.5]$ 中。

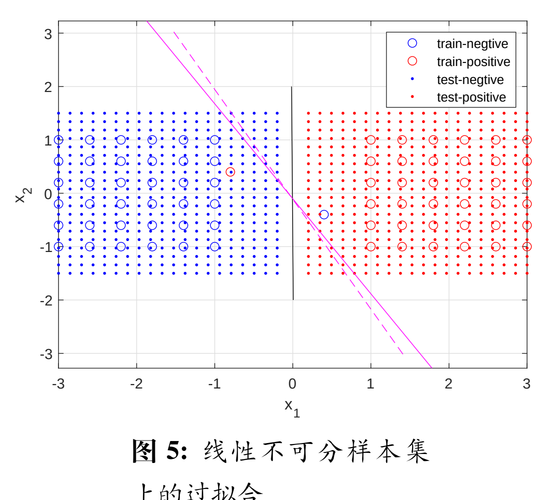


图 5: 线性不可分样本集上的过拟合

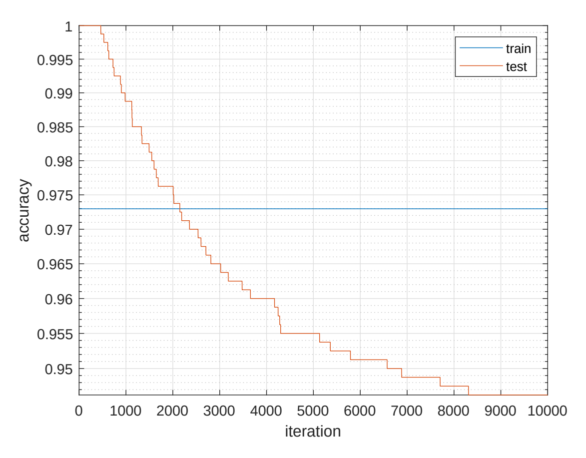


图 6: 线性不可分样本集上的正确率

以交叉熵为损失函数对这个样本集分类，用最速下降法迭代求解式 (3)，迭代步长指定为 0.1。图3中的黑色直线是初始分隔线 (分隔面在二维空间退化为分隔线)，洋红色虚线是迭代 1000 步后的分隔线，洋红色实直线是迭代 2000 步后的分隔线。黑色直线按照式 (9) 选取：

$$(\boldsymbol{x}-\frac{\boldsymbol{\mu}_0+\boldsymbol{\mu}_1}{2})\frac{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T}{|\boldsymbol{\mu}_1-\boldsymbol{\mu}_0|}=0, \tag{9}$$

这里的 $\boldsymbol{\mu}_0$ 是训练集中所有负样本的均值，$\boldsymbol{\mu}_1$ 是训练集中所有正样本的均值。

图4是迭代过程中的正确率走势，在第 1635 步迭代之后，训练集上的正确率达到了 1，但测试集上的正确率从第 180 步开始持续下降，发生过拟合。

4.2 线性不可分样本集上的过拟合

图5中，蓝色圆圈是训练集中的负样本，红色圆圈是训练集中的正样本。训练集中的 36 个负样本均匀分布在区域 $[-3,-1]\times[-1,1]$ 中，一个偏离主体的训练集负样本是点 (0.4,-0.4)。训练集中的 36 个正样本均匀分布在区域 $[1,3]\times[-1,1]$ 中，一个偏离主体的训练集正样本是点 (-0.8,0.4)。根据定义，这个训练集线性不可分。20×20 个蓝色小圆点是测试集中的负样本，它们均匀分布在区间 $[-3,-0.2]\times[-1.5,1.5]$ 中；20×20 个红点小圆点是测试集中的正样本，它们均匀分布在区间 $[0.2,3]\times[-1.5,1.5]$ 中。

以交叉熵为损失函数对这个样本集分类，用最速下降法迭代求解式 (3)，迭代步长指定为 0.1。图5中的黑色直线是初始分隔线，洋红色虚线是迭代 5000 步后的分隔线，洋红色实直线是迭代 10000 步后的分隔线。黑色直线的方程是 (9)。图6是迭代过程中的正确率走势，训练集上的正确率保持平稳，但测试集上的正确率从 463 步开始持续下降，发生过拟合。

仔细观察图3和图5发现，很少的边界样本的大致决定了分隔面的走向，边界样本的影响力比远离边界的样本的影响力大很多，这就是探寻过拟合线索。

4.3 过拟合成因

人眼直观判断，图3和图5中各有 2 个训练样本远离主体，应该按噪音处理，舍去；即使不舍去，它们对确定分隔线的影响也不应太大。实际上，如果舍去噪音样本，那么训练集得到的理

想分隔线应该为 $x_1=0$。黑色直线方程为 $0.9999x_1+0.0140x_2=0$，与理想分割线很接近。

以交叉熵为损失函数得到的分隔线是怎么偏离样本主体的呢？为此，仔细观察损失函数 $h(z,y)$ 的走势。从图2中知道，对正样本 $\boldsymbol{x}_i$，如果 $z_i\geq 0$，那么 $\boldsymbol{x}_i$ 被正确分类，此时它的损失函数值 $h(\sigma(z_i),y_i)\leq -\ln(\sigma(0))$；如果 $z_i<0$，那么 $\boldsymbol{x}_i$ 被错误地分为负类，此时它的损失函数值 $h(\sigma(z_i),y_i)>-\ln(\sigma(0))$。当 $\boldsymbol{x}_i$ 为负样本时，情况类似。

从图2中可以直观地看到，相对于被正确分类的样本，被错误分类的样本对损失函数的贡献更大。

为了定量分析样本对损失函数的贡献，需要用 Taylor 公式寻找 $h(z,y)$ 的简单近似函数。为此定义两个函数

$$f_1(z)=\begin{cases} z-e^z, & \text{如果} z<-C_0<0,\\ \ln(\sigma(z)), & \text{如果}-C_0\leq z\leq C_0,\\ -e^{-z}, & \text{如果} z>C_0>0,\end{cases} \tag{10}$$

$$f_2(z)=\begin{cases} -e^z, & \text{如果} z<-C_0<0,\\ \ln(1-\sigma(z)), & \text{如果}-C_0\leq z\leq C_0,\\ -z-e^{-z}, & \text{如果} z>C_0>0,\end{cases} \tag{11}$$

这里的 C_0 是任意指定的正实数。

定理 1. 函数 $f_1(z)$ 是 $\ln(\sigma(z))$ 的一阶近似，函数 $f_2(z)$ 是 $\ln(1-\sigma(z))$ 的一阶近似。

证. 先证明 $f_1(z)$ 是 $\ln(\sigma(z))$ 的近似。当 $z<-C_0$ 时，$e^z<\exp(-C_0)<1$，从而有

$$\begin{aligned}\ln(\sigma(z)) &= \ln(\frac{1}{1+e^{-z}})=\ln(\frac{e^z}{1+e^z})=\ln(e^z)-\ln(1+e^z)\\ &= z-e^z+O(e^{2z}).\end{aligned}$$

当 $z>C_0$ 时，$e^{-z}<\exp(-C_0)<1$，从而有

$$\ln(\sigma(z))=\ln(\frac{1}{1+e^{-z}})=\ln(1)-\ln(1+e^{-z})=-e^{-z}+O(e^{-2z}).$$

因此，对任意给定的实数 z，$\max(|f_1(z)-\ln(\sigma(z))|)=O(\exp(-2C_0))$，函数 $f_1(z)$ 是 $\ln(\sigma(z))$ 的近似。

再证明 $f_2(z)$ 是 $\ln(1-\sigma(z))$ 的近似。当 $z<-C_0$ 时，$e^z<\exp(-C_0)<1$，从而有

$$\begin{aligned}\ln(1-\sigma(z)) &= \ln(1-\frac{1}{1+e^{-z}})=\ln(1-\frac{e^z}{1+e^z})=\ln(\frac{1}{1+e^z})=\ln(1)-\ln(1+e^z)\\ &= -e^z+O(e^{2z}).\end{aligned}$$

当 $z>C_0$ 时，$e^{-z}<\exp(-C_0)<1$，从而有

$$\begin{aligned}\ln(1-\sigma(z)) &= \ln(1-\frac{1}{1+e^{-z}})=\ln(\frac{e^{-z}}{1+e^{-z}})=\ln(e^{-z})-\ln(1+e^{-z})\\ &= -z-e^{-z}+O(e^{-2z}).\end{aligned}$$

因此，对任意给定的实数 z，$\max(|f_2(z)-\ln(1-\sigma(z))|)=O(\exp(-2C_0))$，函数 $f_2(z)$ 是 $\ln(1-\sigma(z))$ 的近似。 **[证毕]**

当 $C_0=4.3$ 时，$\exp(-C_0)=0.0136, \exp(-2C_0)=0.00018411$。实际上，容易验证，此时有 $0<f_1(z)-\ln(\sigma(z))<0.0001$，$0<f_2(z)-\ln(1-\sigma(z))<0.0001$，逼近良好。

根据定理1，损失函数 (4) 可以近似地表示为

$$h(z,y)\approx\begin{cases}-f_1(z), & \text{如果 } y=1,\\ -f_2(z), & \text{如果 } y=2.\end{cases}$$

为简化说明，本节后续叙述只考虑正样本的损失函数曲线，负样本的情形类似。假设训练过程中的分隔面方程为 $(\boldsymbol{x}-\boldsymbol{c})\boldsymbol{w}=0$，这里的 $\boldsymbol{c}$ 是 d 维行向量。假设样本 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 均为正样本，即 $y_1=1$，$y_2=1$，给定法向量 $\boldsymbol{w}$ 和点 $\boldsymbol{c}$，有 $z_1=(\boldsymbol{x}_1-\boldsymbol{c})\boldsymbol{w}$ 和 $z_2=(\boldsymbol{x}_2-\boldsymbol{c})\boldsymbol{w}$。观察图2中红线知道，如果 $z_1<z_2$，那么 $h(z_1,y_1)>h(z_2,y_2)$，即

推论 1. 样本的加权距离越小，损失贡献越大。

给定 $C_0>0$。当 $z_1<z_2<-C_0$ 时，$\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 均位于分隔面 $(\boldsymbol{x}-\boldsymbol{c})\boldsymbol{w}=0$ 的背面，即都被分错了。假设 $\boldsymbol{w}_3=t\boldsymbol{w}$，其中实数 $t>1$，记 $\bar{z}_1=(\boldsymbol{x}_1-\boldsymbol{c})\boldsymbol{w}_3$ 和 $\bar{z}_2=(\boldsymbol{x}_2-\boldsymbol{c})\boldsymbol{w}_3$，那么有

$$\frac{h(\bar{z}_1,y_1)}{h(\bar{z}_2,y_2)}=\frac{h(tz_1,y_1)}{h(tz_2,y_2)}\approx\frac{-f_1(tz_1,y_1)}{-f_1(tz_2,y_2)}\approx\frac{tz_1-e^{tz_1}}{tz_2-e^{tz_2}}\approx\frac{z_1}{z_2}\approx\frac{h(z_1,y_1)}{h(z_2,y_2)}. \tag{12}$$

由式 (12) 得

推论 2. 被分错样本之间的损失贡献比例不随法向量的变化而变化。

给定 $C_0>0$。当 $C_0<z_1<z_2$ 时，$\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 均位于分隔面 $(\boldsymbol{x}-\boldsymbol{c})\boldsymbol{w}=0$ 的正面，即都被分对了。假设 $\boldsymbol{w}_3=t\boldsymbol{w}$，其中实数 $t>1$，记 $\bar{z}_1=(\boldsymbol{x}_1-\boldsymbol{c})\boldsymbol{w}_3$ 和 $\bar{z}_2=(\boldsymbol{x}_2-\boldsymbol{c})\boldsymbol{w}_3$，那么有

$$\frac{h(\bar{z}_1,y_1)}{h(\bar{z}_2,y_2)}=\frac{h(tz_1,y_1)}{h(tz_2,y_2)}\approx\frac{-f_1(tz_1)}{-f_1(tz_2)}\approx\frac{e^{-tz_1}}{e^{-tz_2}}=\exp(t(z_2-z_1))=(\exp(z_2-z_1))^t, \tag{13}$$

由式 (13) 得

推论 3. 被分对样本之间的损失贡献比例会随着法向量的增长而指数级增长。

给定 $C_0>0$。当 $z_2>C_0$ 且 $z_1=-z_2$ 时，$\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 分别位于分隔面 $(\boldsymbol{x}-\boldsymbol{c})\boldsymbol{w}=0$ 的背面和正面，即一个被分错了另一个被分对了，即 $y_1=2$ 和 $y_2=1$。假设 $\boldsymbol{w}_3=t\boldsymbol{w}$，其中实数 $t>1$，记 $\bar{z}_1=(\boldsymbol{x}_1-\boldsymbol{c})\boldsymbol{w}_3$ 和 $\bar{z}_2=(\boldsymbol{x}_2-\boldsymbol{c})\boldsymbol{w}_3$，那么有

$$\frac{h(\bar{z}_1,y_1)}{h(\bar{z}_2,y_2)}=\frac{h(-tz_2,y_2)}{h(sz_2,y_2)}\approx\frac{-f_1(-tz_2)}{-f_1(tz_2)}\approx\frac{tz_2+\exp(-tz_2)}{\exp(-tz_2)}=1+tz_2\exp(tz_2). \tag{14}$$

由式 (14) 得

推论 4. 被分错样本与被分对样本之间的损失贡献比例会随着法向量的增长而指数级增长。

将分隔面附近样本称为边界样本。从推论1~ 推论4可知，对损失函数的贡献比例，由大到小分顺序是：被分错的样本、被分对的边界样本、被分对的其它样本，它们之间的比例关系随着法向量的增长而迅速增大。适用平面二分类的数据集，被最优分隔面分错的样本占比不大，这样被分错的样本通常会在分隔面附近。由附录定理 2 知，在线性可分数据集上，法向量模长 $|\boldsymbol{w}|$ 趋向无穷大，分隔平面几乎完全由边界样本决定。由附录定理 3 知，在线性不可分数据集上，法向量模长 $|\boldsymbol{w}|$ 有界，但最优分隔面的法向量模长可能仍然很大，过拟合仍然严重。因此得出过拟合原因之一：边界样本的损失贡献比重大且随权重增长而加速增大。

自然界很多事件服从正态分布，例如图7，中心处样本密度大，能够很好在逼近其概率密度函数；在远离中心的边缘处，概率密度函数的值较小，样本稀疏，不能很好地反映其概率密度函数。考虑到训练集边界样本基本决定分隔平面，而测试集样本的实际分布与训练集会有一些

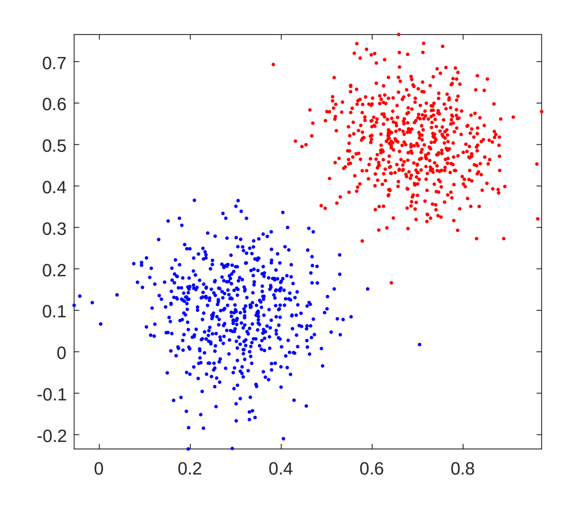


图 7: 一个服从正态分布的样本集

差异，所以得到的分隔平面不能很好地分隔训练集。因此得到过拟合的原因之二：边界样本分布散乱。

第4.1节、第4.2节的 2 个过拟合例子都是根据这 2 个原因设计出来的。

4.4 正则化的作用机理

缓解过拟合的常用手段是添加正则化项，各种各样的正则化方法的目标都是一致的：控制法向量的模长，不让 $|\boldsymbol{w}|$ 过大。由过拟合的成因可知，虽然法向量过大只是过拟合的表象，不是根本原因，但限制它的模长确实有效缓解了过拟合，这是因为它限制了边缘样本的损失贡献比重。正则化缓解过拟的同时，必然会降低训练集上的正确率。

从过拟合成因还可以知道缓解过拟合的另一个思路：修整边界样本使之准确反映概率密度函数。教科书 [14] 中已经写明增加样本数量能缓解过拟合，其实也可以用边界样本散乱的观点来解释：增加样本总量，边界样本数量也同比例增加，从而边界样本更好地反映其概率密度函数，缓解过拟合。

5 附录

先考查损失函数式 (4) 的导数

$$h_z(z,y)=\frac{\partial h(z,y)}{\partial z}=\begin{cases}\sigma(z)-1, & \text{如果 } y=1,\\ \sigma(z), & \text{如果 } y=2.\end{cases} \tag{15}$$

$h_z(z,y)$ 的图像如图8所示。对正样本 $\boldsymbol{x}_i$，当 $z_i>0$ 时 $-0.5<h_z(z_i,y_i)<0$，当 $z_i<0$ 时 $-1<h_z(z_i,y_i)<-0.5$，当 $z_i=0$ 时 $h_z(z_i,y_i)=-0.5$；对负样本 $\boldsymbol{x}_i$，当 $z_i>0$ 时 $0.5<h_z(z_i,y_i)<1$，当 $z_i<0$ 时 $0<h_z(z_i,y_i)<0.5$，当 $z_i=0$ 时 $h_z(z_i,y_i)=0.5$。

定理 2 (**法向无限**). 样本集 D 线性可分时，最优分隔面的法向量 $\hat{\boldsymbol{w}}$ 的模长是 $+\infty$。

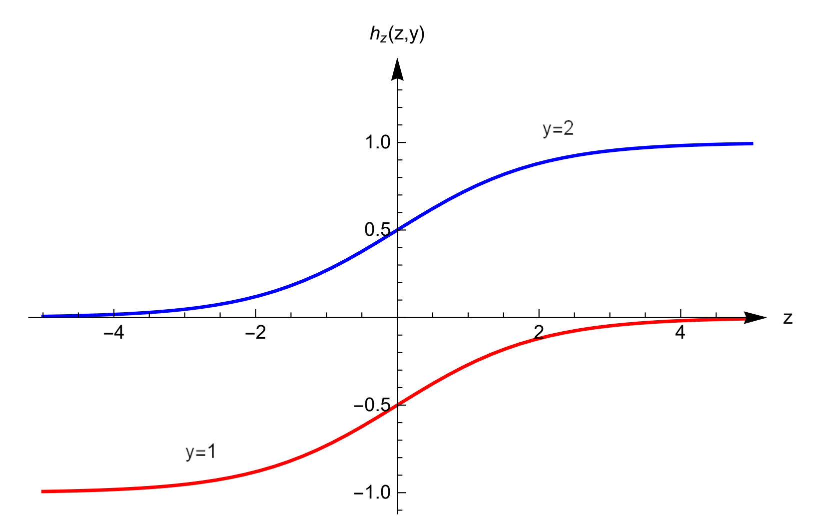


图 8: 损失函数的偏导数 $h_z(z,y)$ 的图像。

证: 根据线性可分的定义，存在列向量 $\tilde{\boldsymbol{w}}_1$ 和行向量 $\tilde{\boldsymbol{c}}$，$|\tilde{\boldsymbol{w}}_1| \neq 0$，例得对 $\forall(\boldsymbol{x}_i, y_i) \in D$ 满足

$$z_i = \begin{cases} (\boldsymbol{x}_i - \tilde{\boldsymbol{c}})\tilde{\boldsymbol{w}}_1 \geq 0, & \text{如果 } y_i = 1, \\ (\boldsymbol{x}_i - \tilde{\boldsymbol{c}})\tilde{\boldsymbol{w}}_1 < 0, & \text{如果 } y_i = 2. \end{cases}$$

令 $\tilde{\boldsymbol{w}}_2 = 2\tilde{\boldsymbol{w}}_1$，那么有

$$\hat{z}_i = (\boldsymbol{x}_i - \tilde{\boldsymbol{c}})\tilde{\boldsymbol{w}}_2 = 2(\boldsymbol{x}_i - \tilde{\boldsymbol{c}})\tilde{\boldsymbol{w}}_1 = 2z_i, i = 1, 2, \ldots, m,$$

$$h(\hat{z}_i, y_i) = \begin{cases} -\ln(1 - \sigma(2z_i)) < -\ln(1 - \sigma(z_i)), & \text{如果 } y_i = 1, \\ -\ln(\sigma(2z_i)) \leq -\ln(\sigma(z_i)), & \text{如果 } y_i = 2. \end{cases}$$

即 $L(\boldsymbol{w}_2, b) < L(\boldsymbol{w}_1, b)$。按照这个每次模长加倍的方法推下去，就得到 $|\hat{\boldsymbol{w}}| = +\infty$。

[证毕]

定理 3 (**法向有界**). 样本集 D 线性不可分时，最优分隔面的法向量有界。

证: 假设满足式 (6) 的最优分隔面的点法式方程为

$$(\boldsymbol{x} - \hat{\boldsymbol{c}})\hat{\boldsymbol{w}} = 0.$$

令 $\boldsymbol{n} = \hat{\boldsymbol{w}}/|\hat{\boldsymbol{w}}|$，显然 $|\boldsymbol{n}| = 1$。假设样本集 D 线性不可分，从而存在指标 i 使得

$$z_i = (\boldsymbol{x}_i - \hat{\boldsymbol{c}})\boldsymbol{n} < 0, \text{且 } y_i = 1, \tag{16}$$

或者

$$z_i = (\boldsymbol{x}_i - \hat{\boldsymbol{c}})\boldsymbol{n} \geq 0, \text{且 } y_i = 2. \tag{17}$$

将指标集合记为

$$\begin{aligned} I_2 &= \{i | z_i < 0 \text{ 且 } y_i = 2, \quad 1 \leq i \leq m\}, \\ I_1 &= \{i | z_i > 0 \text{ 且 } y_i = 1, \quad 1 \leq i \leq m\}, \\ J_2 &= \{j | z_j > 0 \text{ 且 } y_j = 2, \quad 1 \leq j \leq m\}, \\ J_1 &= \{j | z_j < 0 \text{ 且 } y_j = 1, \quad 1 \leq j \leq m\}, \\ K_0 &= \{k | z_k = 0, \quad 1 \leq k \leq m\}, \end{aligned} \tag{18}$$

指标集合上的损失函数分别记为

$$\begin{aligned} L_{I_2}(\boldsymbol{n}) &= \frac{1}{m}\sum_{i\in I_2} h(z_i,y_i), \quad L_{I_1}(\boldsymbol{n}) = \frac{1}{m}\sum_{i\in I_1} h(z_i,y_i), \\ L_{J_2}(\boldsymbol{n}) &= \frac{1}{m}\sum_{j\in J_2} h(z_j,y_j), \quad L_{J_1}(\boldsymbol{n}) = \frac{1}{m}\sum_{j\in J_1} h(z_j,y_j), \\ L_{K_0}(\boldsymbol{n}) &= \frac{1}{m}\sum_{k\in K_0} h(z_k,y_k) \end{aligned} \tag{19}$$

由式 (2) 知，

$$L(\boldsymbol{n},b) = L_{I_2}(\boldsymbol{n}) + L_{I_1}(\boldsymbol{n}) + L_{J_2}(\boldsymbol{n}) + L_{J_1}(\boldsymbol{n}) + L_{K_0}(\boldsymbol{n}).$$

由式 (17)(16) 知 $I_2 \cup I_1 \cup K_0$ 和 $J_2 \cup J_1$ 都是非空集合，为论证方便，这里仅考虑 I_2、I_1、J_2、J_1、K_0 均为非空集合的一般情形，其它特殊情形可做类似证明。

令 $\lambda > 1$ 为正实数，δ 为正无穷小量。接下来寻找 λ 的取值范围，使得

$$\begin{aligned} & L((\lambda+\delta)\boldsymbol{n},b) - L(\lambda\boldsymbol{n},b) \\ = \;& L_{I_2}((\lambda+\delta)\boldsymbol{n}) - L_{I_0}(\lambda\boldsymbol{n}) + L_{I_1}((\lambda+\delta)\boldsymbol{n}) - L_{I_1}(\lambda\boldsymbol{n}) + L_{J_2}((\lambda+\delta)\boldsymbol{n}) \\ & - L_{J_2}(\lambda\boldsymbol{n}) + L_{J_1}((\lambda+\delta)\boldsymbol{n}) - L_{J_1}(\lambda\boldsymbol{n}) + L_{K_0}((\lambda+\delta)\boldsymbol{n}) - L_{K_0}(\lambda\boldsymbol{n}) \\ > \;& 0 \end{aligned} \tag{20}$$

成立。

由式 (18)(7) 知，对 $\forall k \in K_0$，有 $z_k = 0$，$h((\lambda+\delta)z_k,y_k) = h(\lambda z_k,y_k) = \ln(2)$。由式 (19) 知

$$L_{K_0}((\lambda+\delta)\boldsymbol{n}) - L_{K_0}(\lambda\boldsymbol{n}) = 0. \tag{21}$$

对 $\forall j \in J_2$，由式 (4) 知，$h_z(z_j,y_j)$ 的值从 1/2 严格单调递增至 1，从而有

$$\begin{aligned} h((\lambda+\delta)z_j,y_j) - h(\lambda z_j,y_j) > h_z(\lambda z_j,y_j)\delta z_j > \frac{1}{2}\delta z_j, \\ L_{J_2}((\lambda+\delta)\boldsymbol{n}) - L_{J_2}(\lambda\boldsymbol{n}) > \frac{1}{2}\delta\sum_{j\in J_2} z_j. \end{aligned} \tag{22}$$

对 $\forall j \in J_1$，由式 (4) 知，$h_z(z_j,y_j)$ 的值从 -1 严格单调递增至 $-1/2$，从而有

$$\begin{aligned} h((\lambda+\delta)z_j,y_j) - h(\lambda z_j,y_j) > h_z(\lambda z_j,y_j)\delta z_j > -\frac{1}{2}\delta z_j, \\ L_{J_1}((\lambda+\delta)\boldsymbol{n}) - L_{J_1}(\lambda\boldsymbol{n}) > -\frac{1}{2}\delta\sum_{j\in J_1} z_j. \end{aligned} \tag{23}$$

令

$$E_0 = \frac{1}{2}\Big(\sum_{j\in J_2} z_j - \sum_{j\in J_1} z_j\Big)\Big/\Big(-\sum_{i\in I_2} z_i + \sum_{i\in I_1} z_i\Big),$$

由 I_2、I_1、J_2、J_1 的定义知 $E_0 > 0$。这里需要假设 $E_0 < 1$，它可以模糊地理解为“样本集中被分错的样本数量小于被分对的样本数量的 2 倍”，显然是一个合理的假设。由式 (15) 知，在 $y_i = 2$ 时，$h_z(z_i,y_i)$ 在定义域 $z_i \in (-\infty,0)$ 上的值从 0 严格单调递增至 1/2，因此对 $\forall\lambda > \lambda_0 = -\sigma^{-1}(E_0) > 0$ 和 $\forall i \in I_2$ 均有 $h_z(\lambda z_i,y_i) < E_0$。

对 $\forall i \in I_2$，由式 (7)(4) 知

$$0 > h((\lambda+\delta)z_i,y_i) - h(\lambda z_i,y_i) > h_z(\lambda z_i,y_i)\delta z_i,$$

进而，对 $\forall\lambda > \lambda_0$ 有

$$L_{I_2}((\lambda+\delta)\boldsymbol{n}) - L_{I_2}(\lambda\boldsymbol{n}) > \delta\sum_{i\in I_2} h_z(\lambda z_i, y_i)z_i > \delta E_0 \sum_{i\in I_2} z_i. \tag{24}$$

由式 (4) 知，在 $y_i = 1$ 时，$h_z(z_i, y_i)$ 在定义域 $z_i \in (0, +\infty)$ 上的值从 $-1/2$ 严格单调递增至 0，因此对 $\forall\lambda > \lambda_1 = \sigma^{-1}(1-E_0) > 0$ 和 $\forall i \in I_1$ 有

$$-E_0 < h_z(\lambda z_i, y_i) < 0.$$

对 $\forall i \in I_1$，由式 (7)(4) 知

$$0 > h((\lambda+\delta)z_i, y_i) - h(\lambda z_i, y_i) > h_z(\lambda z_i, y_i)\delta z_i,$$

进而，对 $\forall\lambda > \lambda_1$ 有

$$L_{I_1}((\lambda+\delta)\boldsymbol{n}) - L_{I_1}(\lambda\boldsymbol{n}) > \delta\sum_{i\in I_1} h_z(\lambda z_i, y_i)z_i > -\delta E_0 \sum_{i\in I_1} z_i. \tag{25}$$

综合式 (20)-(25) 得知，当 $\lambda > \max\{\lambda_0, \lambda_1\}$ 时，有

$$\begin{aligned}
& L((\lambda+\delta)\boldsymbol{n}, b) - L(\lambda\boldsymbol{n}, b) \\
> \; & \delta E_0 \sum_{i\in I_2} z_i - \delta E_0 \sum_{i\in I_1} z_i + \frac{\delta}{2}\sum_{j\in J_2} z_j - \frac{\delta}{2}\sum_{j\in J_1} z_j + 0 \\
> \; & \delta E_0 (\sum_{i\in I_2} z_i - \sum_{i\in I_1} z_i) + \frac{\delta}{2}(\sum_{j\in J_2} z_j - \sum_{j\in J_1} z_j) \\
\geq \; & \frac{\delta}{2} \frac{\sum_{j\in J_2} z_j - \sum_{j\in J_1} z_j}{-\sum_{i\in I_2} z_i + \sum_{i\in I_1} z_i} (\sum_{i\in I_2} z_i - \sum_{i\in I_1} z_i) + \frac{\delta}{2}(\sum_{j\in J_2} z_j - \sum_{j\in J_1} z_j) \\
= \; & 0.
\end{aligned} \tag{26}$$

式 (26) 意味着 $L(\lambda\boldsymbol{n}, b)$ 在 $\lambda \in (\max\{\lambda_0, \lambda_1\}, +\infty)$ 严格单调递增，从而最分优分隔平面的法向量模长 $|\hat{\boldsymbol{w}}| = \hat{\lambda}|\boldsymbol{n}| = \hat{\lambda} \leq \max\{\lambda_0, \lambda_1\}$，即 $\hat{\boldsymbol{w}}$ 有界。

[证毕]

参考文献

[1] Galen Andrew and Jianfeng Gao. “Scalable Training of L1-Regularized Log-Linear Models”. In: *International Conference on Machine Learning*. 2007. URL: https://www.microsoft.com/en-us/research/publication/scalable-training-of-l1-regularized-log-linear-models/.

[2] Ting Chen et al. “A Simple Framework for Contrastive Learning of Visual Representations”. In: (2020).

[3] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[4] Pedro M. Domingos. “A few useful things to know about machine learning”. In: *Communications of the ACM* 55 (2012), pp. 78 –87.

[5] Pedro M. Domingos. “A Unified Bias-Variance Decomposition and its Applications”. In: 2000.

[6] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. “Regularization paths for generalized linear models via coordinate descent”. In: *Journal of statistical software* 33.1 (2010), pp. 1–22. ISSN: 1548-7660. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929880/ (visited on 04/08/2013).

[7] Chih-Jen Lin and Jorge J. Moré. "Newton's Method for Large Bound-Constrained Optimization Problems". In: *SIAM Journal on Optimization* 9.4 (1999), pp. 1100–1127. DOI: 10.1137/S1052623498345075. eprint: https://doi.org/10.1137/S1052623498345075. URL: https://doi.org/10.1137/S1052623498345075.

[8] Chih-Jen Lin, Ruby Chiu-Hsing Weng, and S. Sathiya Keerthi. "Trust Region Newton Method for Logistic Regression". In: *J. Mach. Learn. Res.* 9 (2008), pp. 627–650.

[9] Simon Perkins and James Theiler. "Online Feature Selection using Grafting". In: *International Conference on Machine Learning*. 2003.

[10] Matthew J. Streeter and H. Brendan McMahan. "Less Regret via Online Conditioning". In: *CoRR* abs/1002.4862 (2010). arXiv: 1002.4862. URL: http://arxiv.org/abs/1002.4862.

[11] R. Tibshirani T. Hastie and J. Friedman. *The Elements of Statistical Learning, Second Edition:Data Mining, Inference, and Prediction. Springer Series in Statistics*. Springer, 2009.

[12] Guo-Xun Yuan, Chia-Hua Ho, and Chih-Jen Lin. "An Improved GLMNET for L1-regularized Logistic Regression". In: *Journal of Machine Learning Research* 13.64 (2012), pp. 1999–2030. URL: http://jmlr.org/papers/v13/yuan12a.html.

[13] Guo-Xun Yuan et al. "A Comparison of Optimization Methods and Software for Large-scale L1-regularized Linear Classification". In: *J. Mach. Learn. Res.* 11 (2010), pp. 3183–3234.

[14] 周志华. 机器学习. 清华大学出版社, Apr. 2016, pp. 57–60.